Sankhyā : The Indian Journal of Statistics
2011, Volume 73-A, Part 2, pp. 267-302

Resampling Methods for the Nonparametric and Generalized Behrens-Fisher Problems

Ansgar Steland
RWTH Aachen University, Aachen, Germany
Appaswamy R. Padmanabhan
Universit Sains Malaysia, Penang, Malaysia
Monash University, Clayton, Australia
Muhammad Akram
Monash University, Clayton, Australia

Abstract

Defining a location parameter as a generalization of the median, a robust test is proposed for (a) the nonparametric Behrens-Fisher problem, where the underlying distributions may have different scales and could be skewed, and (b) the generalized Behrens-Fisher problem, where the distributions may even have different shapes. We propose to bootstrap a signed rank statistic based on differences of sample values and derive rigorous bootstrap central limit theorems for its probabilistic justification, allowing for the so-called m-out-of-n bootstrap. The location parameter of interest is the pseudo-median of the distribution of the difference between a control measurement and an observation from the treatment group. It reduces to (a) the shift in the two sample location model and (b) the difference between the centers of symmetry in the nonparametric Behrens-Fisher model, under the additional assumption that the distributions are symmetric. Due to its importance for applications, we also extend our results to an ANOVA design where each treatment is compared with the control group. Finally, we compare our test with competitors on the basis of theory as well as simulation studies. It turns out that our approach yields a substantial improvement for distributions close to the generalized extreme value type, which makes it attractive for applications in engineering as well as finance. Several heteroscedastic data sets from electrical engineering, astro physics, energy research, analytical chemistry and psychology are used to illustrate our solution.

AMS (2000) *subject classification.* Primary 62G10; Secondary 60F05, 62G35, 62P30.
Keywords and phrases. Bootstrap, central limit theorem, generalized extreme value distribution, heteroscedasticity, signed-rank statistics, U-statistics.

1 Introduction

The classical two sample and multi-sample location problems assume (i) normality, (ii) homogeneity of error variances and (iii) symmetry of the distributions. Those assumptions are not met in many applications, which motivated our work. Particularly, the presence of heteroscedasticity, asymmetry and possibly different

shapes of the distributions causes severe statistical problems. Areas of applications where such issues arise and which we discuss in this article to some extent cover various fields of natural sciences, electrical engineering, photovoltaics as well as psychology and human sciences. The aim of the present article is to propose a new solution to the problem, elaborate on a related bootstrap rank test, provide the required theoretical results, investigate the statistical properties by simulations as well as discuss carefully the application to various real data from the mentioned fields.

The nonparametric Behrens-Fisher framework allows for non-normality and heteroscedasticity, but the rank-based methods studied in Fligner and Policello (1981) and Rust and Fligner (1984) for the two-sample and multi-sample settings, respectively, still assume symmetric distributions. Nevertheless, often of the distributions arising in bioavailability studies, engineering and psychology are skewed (cf. Chow and Liu, 1992, Nair, 1984, Micceri, 1989 and Wilcox, 1987, 1995). Similarly, symmetry can rarely be assumed for financial and economic data. In particular, it is often argued that prices, inflation, returns and other economic measures respond less to *positive news* than to *negative news*. In the economic literature, this is usually referred to as *price stickiness*, *asymmetric responses* or the *ratchet effect*. For further details and references, see Abdus, Ghoudi and Remillard (2003). Such challenging phenomena also arise in photovoltaics, cf. Steland and Zähle (2009) and Herrmann and Steland (2010). In Section 6, we will analyze a data set from that field which is hard to treat with existing methodologies. In psychological experiments a treatment often changes both location and scale of the distribution of the observations, and sometimes even the distributional shape, see Zumbo (2002) and Zumbo and Koh (2005), amongst others. For examples from biometry see Podgor and Gastwirth (1964) and the references given therein. Finally, Boos and Brownie (1991, 2004) and Lamb, Boos and Brownie (1996) show that in toxicological studies it is common that a treatment aspects both location and variability.

These examples demonstrate that the Behrens-Fisher setting is important, both for observational and experimental studies. In real experiments such as pilot studies in sciences and engineering, one often has to base inference on small sample sizes and deals with scientific problems which are not yet well understood. Sometimes, there is simply not enough knowledge about the problem to formulate parametric distributional models or specific alternative models which would allow us to use likelihood-based methods or locally optimal rank tests. In other cases, large samples are not available to conduct such analyzes, e.g. due to cost constraints. For these reasons, investigators frequently complain about classic procedures requiring strong assumptions which can neither be checked nor inferred with available data and theory. Then it is advisable to rely on rank tests using Wilcoxon scores to simultaneously ensure simplicity, robustness, accuracy of type I error, and high power for a wide class of alternative distributions.

Thus, both from a methodological and an applied point of view, there is need for appropriate statistical methods to test for a difference in the presence of asymmetric and heteroscedastic distributions. The present article contributes a novel solution for that classic statistical problem. As we will argue in the next section, the pseudo-median, closely related to the signed Wilcoxon rank sum statistic, provides a promising solution to approach this problem, for which no canonical location parameter exists. It has a meaningful and clear interpretation in the general case of arbitrary (skewed) distributions and reduces to (a) the shift in the two-sample location (shift) model and (b) the difference between the centers of symmetry when the distributions are symmetric, respectively. We propose to base inference on a bootstrap test based on the signed rank statistic calculated from between-sample differences. Our simulations indicate that the resulting procedure provides high power and accurate type I error rates over a wide range of distributions, thus being highly robust with respect to skewness, heteroscedasticity as well as different distributional shapes.

Our test has remarkable strengths for distributions close to the generalized extreme value (GEV) family. Here the Wilcoxon test performs poor both in terms of level and power, whereas our test is highly robust and powerful. Thus, our results provide a reliable testing methodology which is of particular interest in engineering as well as in finance, where such distributions often arise but can not be handled with parametric methods due to the lack of large samples.

Before proceeding, let us briefly discuss some related work. The seminal work of Hodges and Lehmann (1963) on rank tests has been extended in many directions to cover ANOVA and regression models, see Puri and Sen (1971), Denker (1985) and Hettmansperger (1991), among others. However, most results require that, firstly, the distributions have the same shape, and, secondly, the error variances are equal. To test for location and scale effects simultaneously, Podgor and Gastwirth (1964) proposed to use a sum of a rank statistic of location and a rank statistic of scale. However, under the Behrens-Fisher model the statistic has to be modified, since now the scales may differ under the null hypothesis of equal locations. But after these modifications, the null distribution becomes intractable and its power reduces dramatically. In contrast, the method introduced in the present article applies to this problem as well, without such drawbacks.

We shall use the bootstrap to obtain critical values. The bootstrap for the Wilcoxon statistic has been studied in Gill (1989) and for general score statistics by Steland (1998). For a recent study on how to use the bootstrap to quantify the accuracy of a ranking, we refer to Hall and Miller (2009). Choosing the population median as the parameter of interest, a methodology based on the sample median was proposed in Babu, Padmanabhan and Puri (1999), which was applicable not only to the Behrens-Fisher (BF) problem, but also to the generalized Behrens-Fisher (GBF) problem where the underlying distributions may have different shapes. However, this test has typically low power. A more powerful test, which avoids the assumption

of symmetry and is based on the Mann-Whitney statistic, was proposed for the two-sample and multi-sample BF problems in Babu and Padmanabhan (2002) and Babu and Padmanabhan (2007), respectively. Nevertheless, this procedure also has the following drawbacks. Firstly, the consistency of the bootstrap estimator requires the two distributions to have the same shape, rendering it inapplicable to GBF problem. Secondly, the test becomes liberal in unbalanced situations involving relatively small samples with negative pairing. A more detailed account of the drawbacks of these procedures will be given in the Section 5.

The organization of the article is as follows. Section 2 reviews some important aspects of the generalized Behrens-Fisher problem, provides the arguments leading to our proposal and discusses alternative approaches and extensions. Section 3 introduces the proposed signed-rank statistic. We provide the required asymptotic theory, in particular a classic central limit theorem and a bootstrap central limit theorem. Extensions to multi-sample comparisons in an ANOVA design are given in Section 4. Section 5 provides extensive simulation studies which indicate that our proposal works well. Finally, Section 6 illustrates the procedure by analyzing several real data sets from electrical engineering, physics, photovoltaics and psychology. Proofs of the main results and related theoretical results are deferred to an appendix.

2 The Generalized Behrens-Fisher Problem Revisited

This section is devoted to a careful review of some important characteristics of the Behrens-Fisher problem. We provide the arguments leading to our proposal for the problem, the pseudo-median, and discuss its advantages from a methodological point of view. We also show how the approach can be extended to ANOVA designs.

2.1. Modeling aspect. Suppose we want to compare control measurements, e.g. obtained under well defined standard conditions in a laboratory, with treatment measurements where only a part of the treatment conditions can be controlled. This happens frequently, e.g. when comparing stress measurements of a material, say steel, obtained in a lab, with measurements under real conditions where only the amount of stress such as velocity can be controlled, but not the stress due to other factors. Then the shapes of the distributions of the control and treatment group may substantially differ for *any* choice of the (controllable) treatment effect θ. Assuming that the treatment effect θ affects the location of the treatment measurements, we are given a family of distributions $\mathcal{G} = \{G_\theta(\cdot) = G(\cdot - \theta) : \theta \in \mathbb{R}\}$. Here G is a fixed distribution function (d.f.) for the treatment measurements. The real experiment is now given by the pair (F, G_θ) where F denotes the d.f of the control measurements. Fixed treatment conditions correspond to some specific θ, and we will denote the resulting d.f. in the sequel by $G = G_\theta$. For a statistical analysis the statistician has to make precise his or her understanding of the hypothesis of *no difference in location* of the distributions. This is a statistical modeling task. Common choices

are equality of the means,

$$\int x\,dF(x) = \int x\,dG(x),$$

equality of the medians,

$$F^{-1}(0.5) = G^{-1}(0.5),$$

or no stochastic ordering in the sense that

$$p = \int F(x)\,dG(x) = 1/2,$$

but this is not a complete list. Clearly, if we assume a location scale model,

$$F(x) = H((x-\mu_1)/\sigma_1) \quad \text{and} \quad G(x) = H((x-\mu_2)/\sigma_2),$$

for some d.f. H which is symmetric around 0, the above definitions collapse. However, in general this is not the case. Instead, by shifting G, i.e. by picking appropriate values for θ, we may achieve that the means or the medians coincide or that p attains the value $1/2$. This means, in the general setting, the various notions of *no difference in location* imply that different pairs (F, G_θ) represent the null hypothesis. Although, in principle, fixing the experimental or observational conditions for the treatment group fixes a value for θ, the statisticians definition of the *no difference in location* null hypothesis may imply a different one. There is no real solution to this modeling problem, and a careful choice has to be done in each application.

2.2. Unbiasedness as a guide in the dark and the pseudo-median. The aforementioned complications give rise to the adoption of the statistical principle of unbiasedness given a test statistic in the spirit of Bickel and Lehmann (1975). Recall that in the standard one-sample location problem, there is a canonical well defined parameter, namely the center of symmetry. It coincides with the expectation of all important estimators including R-estimators, M-estimators and L-estimators. This also eases (asymptotic) comparisons of these estimators and derived tests, which are now mainly based on comparisons of the (asymptotic) variances. If we drop the assumption of symmetry, each of these estimators has a different expectation. Thus, there is no unique measure of location. Bickel and Lehmann (1975) argued along the following lines.

> *"Given an estimator T, the only parameter of interest is its expectation $E(T)$. Therefore inference should be based on that T, which estimates $E(T)$ satisfactorily for a wide class of distributions. In other words, that parameter should be chosen which leads to a satisfactory solution".* (See Bickel and Lehmann, 1975, Sections 5 and 6).

Further, if $E(T)$ depends on the sample size n, the parameter of interest is $\lim_{n\to\infty} E(T)$.

We adopt this viewpoint and take the Wilcoxon signed rank statistic as a starting point. Recall that its asymptotic expectation is related to the parameter (functional) $P(Z_1 > -Z_2)$, when applied to an i.i.d. sample $Z_1, \ldots, Z_n$ with a common continuous distribution function. This fact suggests to consider the parameter $P(Z_1 > -Z_2)$ and related parameters, respectively. Clearly, the Z_i's may be some function of the original data, and we will now argue how to choose them to obtain a parameter yielding a satisfactory solution with a corresponding test statistic of the signed rank sum type.

Assume we are given two independent samples $X, X_1, \ldots, X_n \sim F(x)$ and $Y, Y_1, \ldots, Y_m \sim G(x)$ of i.i.d. observations. Let us also assume for a moment that $\mathrm{Var}\,(X) = \mathrm{Var}\,(Y)$ exists and is known. Denote the corresponding means by $\mu_X = E(X)$ and $\mu_Y = E(Y)$, respectively. The classical Gauss test considers the difference of the means, $\mu_X - \mu_Y$, as the treatment effect of interest and bases inference on the test statistic

$$T_{n,m} = \sqrt{nm/(m+n)}(\overline{X}_n - \overline{Y}_m),$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\overline{Y}_m = m^{-1}\sum_{i=1}^m Y_i$ are the minimum variance unbiased estimators. The central limit theorem (CLT) ensures that

$$T_{n,m} - \sqrt{nm/(m+n)}(\mu_X - \mu_Y)$$

converges in distribution to a normal distribution with median 0. This means, for large sample sizes the distribution of $T_{n,m}$ can be approximated by a normal distribution with mean (=median) $\sqrt{nm/(m+n)}(\mu_X - \mu_Y)$. If the observations are normal, that median is given by

$$\delta_{n,m} = \mathrm{med}(T_{n,m}).$$

This fact suggests to consider the median $\delta_{n,m}$ for finite sample sizes as a reasonable treatment effect, whether or not the data are normally distributed. The smallest sample sizes yielding a non-trivial averaging procedure are $n = m = 2$. In this case, we obtain

$$\delta = \delta_{2,2} = \mathrm{med}(\overline{X}_2 - \overline{Y}_2).$$

Put

$$D_{ij} = X_i - Y_j, \qquad i = 1, \ldots, n,\ j = 1, \ldots, m,$$

and note that

$$\delta = \frac{1}{2}\mathrm{Med}(X_1 + X_2 - Y_1 - Y_2) = \frac{1}{2}\mathrm{Med}(D_{11} + D_{22}). \tag{2.1}$$

δ is called *pseudo-median* of the distribution of $D_{11} \stackrel{d}{=} D_{ij}$, such that $\delta > 0$ indicates that there is a positive location shift of the X-sample compared to the Y-sample. In what follows, we shall take (2.1) as the definition of δ. Note that δ coincides with the shift parameter in the two-sample location problem, and reduces to the

difference between the centers of symmetry in the BF problem, under the additional assumption that the distributions are symmetric.

The associated two-sided testing problem is given by

$$H_0 : \delta = 0 \qquad \text{versus} \qquad H_1 : \delta \neq 0. \tag{2.2}$$

Assuming that $D_{11} + D_{22}$ has an unique median, we have

$$P(D_{11} + D_{22} \leq 2\delta) = P(D_{11} + D_{22} \geq 2\delta) = 1/2$$

and the testing problem (2.2) naturally transforms to

$$H_0 : \theta(F, G) = 1/2 \qquad \text{versus} \qquad H_1 : \theta(F, G) \neq 1/2, \tag{2.3}$$

where $\theta(F, G) = P(D_{11} + D_{22} > 0)$ is the probability being related to the singed rank sum statistic when applied to between-sample differences. To the best of our knowledge, these interesting relationships between the parameters $\delta_{n,m}$, δ and $\theta(F, G)$ have not yet been discussed in the literature.

It is worth mentioning that an interesting class of alternative distributions for the d.f.

$$F_D(x) = F_D^{(F,G)}(x) = \int F(x + y)\, dG(y)$$

of $D_{11} = X_1 - Y_1$ is the class of absolutely continuous and stochastically positive d.f.s H, i.e., $H(x) + H(-x) \leq 1$ for all x with strict inequality for an interval of x values. If $F_D^{(F,G)}$ belongs to that class, large differences are more frequently observed than small differences in the sense that for any $x > 0$

$$P(D_{11} \geq x) \geq P(D_{11} \leq -x)$$

with strict inequality for an interval of x values, and $\theta(F, G) > 1/2$, see Hettmansperger (1991, p. 49).

2.3. Generalizations to ANOVA designs. Notice that the formulation of hypotheses in terms of the pseudo-median can be easily generalized to the important one-factorial ANOVA problem to compare $a - 1$ treatments with a control.

Here, we are given a independent samples $X_{k1}, \ldots, X_{kn_k} \sim F_k(x)$ with sample sizes n_k, $k = 1, \ldots, a$. Assume that $k = 1$ corresponds to the control group. Denote $D_{klij} = X_{ki} - X_{lj}$, $i = 1, \ldots, n_k$, $j = 1, \ldots, n_l$. The testing problem of interest can be formulated as follows. Define $\mathcal{H}_0 = \{(1, 2), \ldots, (1, a)\}$. We aim at testing the null hypothesis

$$H_0 : \text{Med}(D_{kl11} + D_{kl22}) = 0 \text{ for all } (k, l) \in \mathcal{H}_0 \tag{2.4}$$

against

$$H_1 : \text{Med}(D_{kl11} + D_{kl22}) \neq 0 \text{ for some pair } (k, l) \in \mathcal{H}_0\ . \tag{2.5}$$

Notice that when there are many treatments, it is reasonable to select only certain treatments of interest to increase power.

2.4. Remarks on alternative parameters. In Section 5, we will compare our procedure with some competitors taking into account theoretical considerations and simulation results. At this point, let us close this section with a brief comparison with another candidate parameter, the median, $\mathrm{Med}(X_1 - Y_1)$, of the distribution of $X_1 - Y_1$.

Unfortunately, for that parameter a nice theory for the BF and GBF models does not exist. The only method which works in the BF setting without the symmetry assumption is the procedure due to Babu and Padmanabhan (2002). It may be possible to modify it to conduct inference for $\mathrm{Med}(X_1 - Y_1)$. But such a modification would inherit the weaknesses of the approach. Particularly, it would yield liberal tests in unbalanced designs involving relatively small samples with negative pairing.

On the contrary, the pseudo-median has some compelling advantages. Firstly, as shown in the next section, it can be consistently estimated by a Hodges-Lehmann type estimator based on a random sample. Further, testing problems formulated in terms of the pseudo-median can be easily tested using the proposed bootstrap test. In this sense, it leads to an unifying applicable approach for estimation and testing. Secondly, as discussed above, it collapses to known measures of location if the data satisfy the more restrictive classic assumptions. Furthermore, the test statistic even may become distribution-free. For practical applications, particularly for small samples, these properties are very beneficial: The investigator can analyze his or her data under extremely weak assumptions having the guarantee that both the testing problem and the test statistic become classical, provided the data satisfy the classical assumptions. Thirdly, it provides a rather natural generalization of the Hodges and Lehmann (1963) approach to the BF and GBF models. Finally, it can be extended to ANOVA designs, and, presumably, to regression designs as well.

3 Test Statistics and Bootstrap Test for the Two-sample Problem

To deal with the testing problem (2.2), we use a signed rank test applied to the differences between the samples. It turns out that the asymptotic distribution of the test statistic depends on unknown parameters which are difficult to estimate, at least in small samples. Thus, we propose to apply an appropriate bootstrap scheme to obtain critical values. However, the known results about the bootstrap of simple linear rank statistics due to Gill (1989) and Steland (1997, 1998), are not directly applicable to our problem and the same applies to bootstrap central limit theorems for U-statistics, cf. the discussion in Subsection 3.3. and the appendix, which provides a detailed study of the asymptotic theory leading to our results. For general reviews of the bootstrap we refer to Babu and Rao (1993) and Shao and Yu (1995).

3.1. Signed rank test for the differences and its asymptotic distribution. Let us assume that the random variables X_{ij}, $1 \le j \le n_i$, $i = 1, 2$, are defined on a common probability space $(\Omega, \mathcal{A}, P)$. Recall that $D_{ij} = X_i - Y_j$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, are the $N = nm$ differences between the samples and denote by R_{ij}

the rank of $|D_{ij}|$ among $|D_{kl}|$, $k = 1, \ldots, n$, $l = 1, \ldots, m$, defined as

$$R_{ij} = N\widehat{H}_N(|D_{ij}|),$$

where

$$\widehat{H}_N(x) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{1}(|D_{ij}| \le x), \qquad x \in \mathbb{R}.$$

Now the signed rank statistic is given by

$$W_N = \sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\mathbf{1}(D_{ij} > 0) - \frac{N(N+1)}{4}.$$

If $\delta > 0$, we expect to observe more positive (large) differences $D_{ij} = X_i - Y_j$ leading to a large value of the statistic W_N. We shall also consider the scaled version $T_N = \frac{2\sqrt{n+m}}{N(N+1)}W_N$, i.e.

$$T_N = \sqrt{n+m}\left\{\frac{2}{N(N+1)}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\mathbf{1}(D_{ij} > 0) - \frac{1}{2}\right\}.$$

The following theorem provides the asymptotic distribution of the signed rank statistic T_N applied to the mn differences under general fixed alternatives, which is of some interest in its own right. The asymptotic distribution of the statistic T_N under the null hypothesis is covered as a special case. The proof is deferred to the appendix. In what follows, we shall signify convergence in distribution by $\xrightarrow{d}$.

THEOREM 3.1. *Suppose $n/(n+m) \to \lambda \in (0,1)$, as $n, m \to \infty$. Then, for arbitrary fixed distribution functions F and G,*

$$\sqrt{n+m}\left\{\binom{N}{2}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\mathbf{1}(D_{ij} > 0) - \theta(F,G)\right\} \xrightarrow{d} N(0, \eta^2(F,G)),$$

as $n, m \to \infty$, where the asymptotic variance $\eta^2(F,G)$ is given by

$$\eta^2(F,G) = 4\lambda^{-1}\sigma_{01}(F,G) + 4(1-\lambda)^{-1}\sigma_{10}(F,G) \tag{3.1}$$

with

$$\begin{aligned}
\sigma_{01}(F,G) &= \int\left[\int(1 - F(y_1 - x - y_2))\,dG(y_1)\,dG(y_2)\right]dF(x),\\
\sigma_{10}(F,G) &= \int\left[\int(1 - F(y - x_1 + y_2))\,dF(x_1)\,dG(y_2)\right]dG(y).
\end{aligned}$$

Let us now consider the Hodges Lehmann estimator related to the signed rank statistic to estimate the parameter $\delta = (1/2)\mathrm{Med}(D_{11} + D_{22})$ quantifying the difference in location between the X- and the Y-sample. For the present setting, it is given by

$$\widehat{\delta}_N = \widehat{\mathrm{Med}}(D_{ij} + D_{kl})/2,$$

where here and in the sequel $\widehat{\mathrm{Med}}(\xi_j)$ denotes the sample median of a sample ξ_j. This means, $\widehat{\delta}_N$ is obtained by calculating the sample median of the N^2 sums

$$D_{ij} + D_{kl}, \quad i,k = 1,\dots,n,\ j,l = 1,\dots,m.$$

The following theorem, a strong law of large numbers, yields the consistency of $\widehat{\delta}_N$.

THEOREM 3.2. *Suppose that the d.f. of $D_{11} + D_{22}$ is continuous at $1/2$. Then the estimator $\widehat{\delta}_N$ is strongly consistent for the pseudo-median δ, i.e.,*

$$\widehat{\delta}_N \overset{P-a.s.}{\to} \delta,$$

as $n, m \to \infty$.

3.2. Bootstrap procedure. We propose to use a bootstrap procedure to obtain critical values for the test statistics, in order to construct hypothesis tests. To simplify the exposition of the bootstrap algorithm, let us denote the empirical distribution function of a given sample $\xi_1,\dots,\xi_l$ by e.d.f.$(\xi_1,\dots,\xi_l)$.

BOOTSTRAP ALGORITHM:

1. The samples $X_1,\dots,X_n$ and $Y_1 + \widehat{\delta}_N,\dots,Y_m + \widehat{\delta}_N$ mimic a H_0-sample with no difference in location. Define the corresponding differences

   $$\widehat{D}_{ij} = X_i - Y_j - \widehat{\delta}_N, \qquad i = 1,\dots,n,\ j = 1,\dots,m,$$

   which mimic the differences under H_0.

2. Select bootstrap sample sizes $n^*, m^* \in \mathbb{N}$ and resample with replacement bootstrap samples $X_1^*,\dots,X_{n^*}^*$ from $X_1,\dots,X_n$ and $Y_1^*,\dots,Y_{m^*}^*$ from $Y_1 + \widehat{\delta}_N,\dots,Y_m + \widehat{\delta}_N$, i.e.,

   $$X_1^*,\dots,X_{n^*}^* \overset{i.i.d.}{\sim} \text{e.d.f}(X_1,\dots,X_n) \tag{3.2}$$

   $$Y_1^*,\dots,Y_{m^*}^* \overset{i.i.d.}{\sim} \text{e.d.f}(Y_1 + \widehat{\delta}_N,\dots,Y_m + \widehat{\delta}_N). \tag{3.3}$$

   Note that the bootstrap sample sizes n^* and m^* may differ from n and m. Put $N^* = n^* + m^*$.

3. Define the bootstrap sample of the $N^* = n^* m^*$ differences,

   $$D_{ij}^* = X_i^* - Y_j^*, \qquad i = 1,\dots,n^*,\ j = 1,\dots,m^*. \tag{3.4}$$

Then the bootstrap version $T^*_{N^*}$ of T_N is given by

$$T^*_{N^*} = \sqrt{n^* + m^*}(\widetilde{W}^*_{N^*} - \widehat{C}_N), \tag{3.5}$$

where

$$\widetilde{W}^*_{N^*} = \frac{2}{N^*(N^*+1)} \sum_{i=1}^{n^*} \sum_{j=1}^{m^*} R^*_{ij} \mathbf{1}(D^*_{ij} > 0),$$

$$\widehat{C}_N = \frac{2}{N(N+1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \widehat{R}_{ij} \mathbf{1}(\widehat{D}_{ij} > 0).$$

Here R^*_{ij} denotes the rank of $|D^*_{ij}|$ among $|D^*_{kl}|$, $k = 1, \ldots, n^*$, $l = 1, \ldots, m^*$, and $\widehat{R}_{ij}$ is the rank of $|\widehat{D}_{ij}|$ among $|\widehat{D}_{kl}|$, $k = 1, \ldots, n$, $l = 1, \ldots, m$. Notice that $W^*_{N^*} = \frac{N^*(N^*+1)}{2\sqrt{n^*+m^*}} T^*_{N^*}$.

BOOTSTRAP TESTS:

The construction of a bootstrap test with nominal type I error rate $\alpha \in (0, 1)$ can now be based on either W_N or T_N. We describe the procedure for the statistic W_N, since the treatment of T_N is then straightforward. Repeat the bootstrap B times to obtain a sample $W^*_{N^*}(b)$, $b = 1, \ldots, B$, of bootstrap replicates of $W^*_{N^*}$, the bootstrap version of W_N. Denote by $q^*_B(p)$ the empirical *pth* quantile of $W^*_{N^*}(b)$, $b = 1, \ldots, B$. The null hypothesis $H_0 : \delta \geq 0$ is rejected in favor of $H_1 : \delta < 0$, if $W_N < q^*_B(\alpha)$. Similarly, $H_0 : \delta \leq 0$ is rejected in favor of $H_1 : \delta > 0$, if $W_N > q^*_B(1-\alpha)$. Finally, we propose to reject the null hypothesis $H_0 : \delta = 0$ in favor of $H_1 : \delta \neq 0$, if $|W_N| > |q^*_B|(1-\alpha/2)$, where $|q^*_B|(p)$ denotes the p-quantile of the boostrap distribution of W^*_N. Indeed, we obtained better results in our simulation study for this test than for the combined two-tailed rule which rejects H_0, if $W_N < q^*_B(\alpha/2)$ or $W_N > q^*_B(1-\alpha/2)$.

3.3. Bootstrap central limit theorem. The following bootstrap central limit theorem provides the strong consistency of the bootstrap distribution estimator of the test statistic in the sense that

$$\sup_{x \in \mathbb{R}} |P^*(T^*_{N^*} \leq x) - P(T_N \leq x)| \to 0, \qquad P\text{-a.s.},$$

as $n, m \to \infty$ and $n^*, m^* \to \infty$ under natural conditions on the (bootstrap) sample sizes. Recall that the result follows, if, under the bootstrap probability P^*, the sequence of bootstrap statistics $T^*_{N^*}$ converges in distribution to the same law, P-almost surely, as the original statistic T_N under the probability measure P.

Due to the specific and new resampling scheme where one resamples from original X_i's and estimated residuals $Y_i + \widehat{\delta}_N$, we provide a detailed proof in the appendix. The proof makes use of the fact that we may approximate the signed rank statistic by an appropriate U-statistic both under the probability measure P as well as

under the (conditional) bootstrap probability measure P^*. But then the term used to center the bootstrap version, the conditional expectation of the U-statistic under P^*, differs from the term we want to use, namely the signed rank statistic calculated from estimated residuals. For these reasons, known bootstrap central limit theorems for U-statistics are not directly applicable to our problem.

THEOREM 3.3. *Under the conditional bootstrap probability measure P^* induced by the bootstrap scheme, we have P-a.s.*

$$T^*_{N^*} \xrightarrow{d} N(0, \eta^2(F,G)),$$

with $\eta^2(F,G)$ given by (3.1), if $N, N^* \to \infty$ such that for some $\lambda \in (0,1)$

$$n^*/(n^*+m^*) \to \lambda, \quad \text{and} \quad n/(n+m) \to \lambda.$$

We close this section with some remarks.

REMARK 3.1. The extensions to average ranks are straightforward and obtained by substituting the indicator $\mathbf{1}(\cdot \geq 0)$ by its normalization, $\mathbf{1}(\cdot > 0) + \frac{1}{2}\mathbf{1}(\cdot = 0)$, in the above definitions. Also notice that the corresponding operator defines an isometric isomorphism between the Skorohod space $D[0,1]$ and the space of normalized cadlag functions, see Steland (1998) for details.

REMARK 3.2. Our bootstrap scheme as well as the bootstrap central limit theorem allow for the m-out-of-n bootstrap. However, a study to which extent one may improve the bootstrap approximation by choosing $(n^*, m^*) \neq (n, m)$ is beyond the scope of the present article.

4 Bootstrap Test for Many-to-one Comparisons

The two-sample bootstrap rank test can be easily generalized to the many-to-one comparison problem assuming a one-factorial ANOVA design. Suppose we are given a independent samples, $a-1$ treatment groups and a control group. Suppose that $\mathcal{H}_0 \subset \{(1,2), \ldots, (1,a)\}$. To test the null hypothesis (2.4) against the alternative (2.5) one may proceed as follows.

4.1. Testing many-to-one comparison. To the k-th and l-th sample we may associate the $N_{kl} = n_k n_l$ differences $D_{klij} = X_{ki} - X_{lj}$, $i = 1, \ldots, n_k$, $j = 1, \ldots, n_l$, and define the test statistic

$$T_{kl} = \sqrt{n_k + n_l} \left\{ \left(\frac{N_{kl}(N_{kl}+1)}{2} \right)^{-1} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \mathbf{1}(D_{klij} > 0) R_{ij}^{(k,l)} - \frac{1}{2} \right\},$$

where $R_{ij}^{(k,l)}$ denotes the rank of $|X_{ki} - X_{lj}|$ in the combined sample $|X_{ki'} - X_{lj'}|$, $i' = 1, \ldots, n_k$, $j' = 1, \ldots, n_l$.

To test the null hypothesis H_0 one may use the test statistic

$$T_{\mathcal{H}_0} = \sum_{(k,l) \in \mathcal{H}_0} |T_{kl}|.$$

H_0 is rejected for large values of $T_{\mathcal{H}_0}$. We will show that $T_{\mathcal{H}_0}$ converges in distribution.

4.2. Bootstrap test. The bootstrap works by resampling each statistic T_{kl} as in the two-sample situation described above. Noting that for each $(k,l) \in \mathcal{H}_0$ we have $k = 1$, let

$$\widehat{\delta}_{1l} = \widehat{\text{Med}}(D_{1lij} + D_{1lrs})/2$$

where the empirical median is calculated from the $N_1 N_l$ terms $D_{1lij} + D_{1lrs}$, $i, r = 1, \ldots, n_1$, $j, s = 1, \ldots, n_l$. Define

$$\widehat{D}_{1lij} = X_{1i} - X_{lj} - \widehat{\delta}_{1l},$$

for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_l$.

Draw nonparametric bootstrap samples $X^*_{11}, \ldots, X^*_{1n_1^*}$ from the sample $X_{11}, \ldots, X_{1n_1}$, and $X^*_{l1}, \ldots, X^*_{ln_l^*}$ from $X_{l1} + \widehat{\delta}_{1l}, \ldots, X_{ln_l} + \widehat{\delta}_{1l}$ such that

$$X^*_{11}, \ldots, X^*_{1n_1^*} \overset{i.i.d.}{\sim} \text{e.d.f.}(X_{11}, \ldots, X_{1n_1}),$$

$$X^*_{l1}, \ldots, X^*_{ln_l^*} \overset{i.i.d.}{\sim} \text{e.d.f.}(X_{l1} + \widehat{\delta}_{1l}, \ldots, X_{ln_l} + \widehat{\delta}_{1l}).$$

Define

$$D^*_{1lij} = X^*_{1i} - X^*_{lj}$$

for $i = 1, \ldots, n_1^*$ and $j = 1, \ldots, n_l^*$. Put $N^*_{1l} = n_1^* + n_l^*$. The bootstrap version of T_{1l} is then given by

$$T^*_{1l} = \sqrt{n_1^* + n_l^*} \left\{ \binom{N^*_{1l}}{2}^{-1} \sum_{i=1}^{n_1^*} \sum_{j=1}^{n_l^*} \mathbf{1}(D^*_{1lij} > 0) R^{(1,l)*}_{ij} - \binom{N_{1l}}{2}^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_l} \mathbf{1}(\widehat{D}_{1lij} > 0) \widehat{R}^{(1,l)}_{ij} \right\} \tag{4.1}$$

where $R^{(1,l)*}_{ij}$ denotes the rank of $|D^*_{1lij}|$ among $|D^*_{1li'j'}|$, $i' = 1, \ldots, n_1^*$, $j' = 1, \ldots, n_l^*$, and $\widehat{R}^{(1,l)}_{ij}$ denotes the rank of $|\widehat{D}_{1lij}|$ among $|\widehat{D}_{1li'j'}|$, $i' = 1, \ldots, n_1$, $j' = 1, \ldots, n_l$. The bootstrap version of $T_{\mathcal{H}_0}$ is now given by

$$T^*_{\mathcal{H}_0} = \sum_{(k,l) \in \mathcal{H}_0} |T^*_{kl}|.$$

Repeat the resampling step B times to obtain a sample $T^*_{\mathcal{H}_0}(b)$, $b = 1, \ldots, B$, of bootstrap replicates. Now denote by $|q|^*_B(p)$ the empirical p-quantile of $|T^*_{\mathcal{H}_0}(b)|$, $b = 1, \ldots, B$. Then the two-sided bootstrap rank test for the testing problem (2.4) rejects H_0 if $|T_{\mathcal{H}_0}| > |q|^*_B(1 - \alpha)$.

4.3. Bootstrap central limit theorem. The asymptotic justification of the bootstrap procedure is based on the following multivariate central limit theorem and the associated bootstrap central limit theorem for the random vector $(T_{kl})_{(k,l)\in\mathcal{H}_0}$ on which our tests statistic $T_{\mathcal{H}_0}$ is based.

THEOREM 4.1. *Under the null hypothesis (2.4), the following assertions hold true.*

(i) We have

$$(T_{kl})_{(k,l)\in\mathcal{H}_0} \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{H}_0}), \quad \mathbf{\Sigma}_{\mathcal{H}_0} = diag\,(\eta^2(F_k, F_l) : (k,l) \in \mathcal{H}_0),$$

provided $n_1, \ldots, n_a \to \infty$ *with* $n_k/(n_k+n_l) \to \lambda_{kl} \in (0,1)$, *for all* $(k,l) \in \mathcal{H}_0$.

(ii) Under the conditional bootstrap probability measure P^*, *we have* P*-a.s.*

$$(T^*_{kl})_{(k,l)\in\mathcal{H}_0} \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{H}_0}),$$

provided $n_1, \ldots, n_a \to \infty$ *and* $n^*_1, \ldots, n^*_a \to \infty$ *such that* $n_k/(n_k+n_l) \to \lambda_{kl} \in (0,1)$ *and* $n^*_k/(n^*_k+n^*_l) \to \lambda_{kl}$ *for all* $(k,l) \in \mathcal{H}_0$.

By virtue of the continuous mapping theorem, we obtain the following corollary concerning the statistics $T_{\mathcal{H}_0}$ and $T^*_{\mathcal{H}_0}$, which are continuous functions of $(T_{kl} : (k,l) \in \mathcal{H}_0)$ and $(T^*_{kl} : (k,l) \in \mathcal{H}_0)$, respectively.

COROLLARY 4.1. *Under the assumptions of Theorem 4.1 the following assertions hold true.*

(i) We have

$$T_{\mathcal{H}_0} \xrightarrow{d} \sum_{(k,l)\in\mathcal{H}_0} |Z_{kl}|,$$

where $(Z_{kl} : (k,l) \in \mathcal{H}_0) \sim N(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{H}_0})$, *if* $n_1, \ldots, n_a \to \infty$ *such that* $n_k/(n_k+n_l) \to \lambda_{kl} \in (0,1)$, *for all* $(k,l) \in \mathcal{H}_0$.

(ii) Under the conditional bootstrap scheme we have P*-a.s.*

$$T^*_{\mathcal{H}_0} \xrightarrow{d} \sum_{(k,l)\in\mathcal{H}_0} |Z_{kl}|,$$

provided $n_1, \ldots, n_a \to \infty$ *and* $n^*_1, \ldots, n^*_a \to \infty$ *such that* $n_k/(n_k+n_l) \to \lambda_{kl} \in (0,1)$ *and* $n^*_k/(n^*_k+n^*_l) \to \lambda_{kl}$ *for all* $(k,l) \in \mathcal{H}_0$.

5 Comparisons and Simulations

This section is devoted to a detailed comparison of the proposed method with competitors by comparing their theoretical properties and investigating them by

Monte Carlo simulation. Both the two-sample setting and the many-to-one ANOVA design are studied with a focus on the former. Our Monte Carlo study covers various common simulation models as well as some members of the generalized extreme value (GEV) family of distributions. Indeed, it turns out that for GEV distributions our test has exceptional properties in terms of level and power and outperforms the classic Wilcoxon test, although the test is still easy to apply. Moreover, there is no price to pay, since for other distributions such as mixtures of normals our test is at least as good and often better than its competitors.

We start with a careful comparison of the proposed approach with some competitors, which takes into account both theoretical considerations and simulation results from earlier studies. As a result, we can confine our Monte Carlo simulations considerably. Those simulations focus on the case of small samples in the presence of unbalanced sample sizes, heteroscedasticity and asymmetry. We simulated the type I error rate as well as the power under a location shift under various distributions corresponding to the above phenomena. The simulations were done in R using C functions to speed up calculations.

5.1. Comparison with some alternative procedures. The competitors to our approach are the procedures of Babu-Padmanabhan-Puri, the classic Mann-Whitney-Wilcoxon test, the approach of Fligner-Policello and the Babu-Padmanabhan (BP) procedures. We explain their drawbacks and also how our procedure is free of those drawbacks.

Let us start with the test of Babu, Padmanabhan and Puri (1999). This test has typically low powers. In the two sample case, it is equivalent to the median test, whose asymptotic relative efficiency (ARE) in the normal model is only 0.637 (c.f. Lehmann, 1975, p. 379). Next consider the Mann-Whitney-Wilcoxon statistic U, the number of pairs (X_i, Y_i) with $X_i < Y_i$. Denote its null variance, lower, and upper quantiles in the location model by $V(L), q(L)$, and $Q(L)$, respectively. Further, denote by $V(BF), q(BF)$, and $Q(BF)$ their counterparts in the BF model. Recall that $V(L) = mn(m+n+1)/12$ (c.f. Lehmann, 1975, p. 14). Let us consider the BF model. First assume that F and G are symmetric. In this case, $V(BF)$ will be typically greater than $V(L)$ and also unknown. As a result, $q(BF)$ and $Q(BF)$ will be far more dispersed than $q(L)$ and $Q(L)$, respectively, yielding a liberal test. Increasing the sample even makes things worse. The larger the sample size, the greater the difference between $q(L)$ and $q(BF)$ (and similarly between $Q(L)$ and $Q(BF)$). This is also borne out by the results in Table 2 of Fligner and Policello (1981, p. 166).

Let τ be the ratio of the scale parameters. The standard location model corresponds to the case $\tau = 1$. In this case U is robust. However, in the BF model for $m = 25$ and $n = 20$, the empirical level of the corresponding test is 8.9% (when $\tau = 10$) even for the normal distribution. For skewed distributions, the performance will be even worse, as will be explained shortly. The Fligner-Policello (FP) procedure seeks to overcome this drawbacks by obtaining a variance estimate, say v, being consistent under their assumptions, and uses the statistic $(U/mn - 0.5)/\sqrt{v}$

(formula 3.2, p164), which yields good results in the symmetric case. Next consider the skewed case and assume, for simplicity, the parameter of interest is the population median. Then the null hypothesis H_0 is that the population medians are equal and, without loss of generality, can be assumed to be zero. Let $p = P_0(X_1 < Y_1)$, where P_0 indicates that the probability is calculated under H_0. In this setting, the variance estimate of the FP procedure is still correct, but the mean is wrong, since only under symmetry we have $p = 0.5$. In the skewed case, p will be unknown and typically different from 0.5. As will be explained in the next paragraph, p could be close to 0 or close to 1. Therefore the FP procedure centers U/mn at the wrong value, 0.5, instead of the much smaller or much larger value p. As a result, its limiting normal distribution will have the correct variance but the wrong mean, making the test liberal for one tail and conservative for the other.

Let ρ denote the ratio of the scale parameter of G to that of F. Suppose F and G have the same shape, the same median 0 and are right-skewed. Now it can be shown that, if $\rho > 1$, then $p < 0.5$ and as the skewness increases, p will decrease and move towards zero (similarly if $\rho < 1$, p will move towards one). Suppose $\rho > 1$. If F and G are exponential, then it is known that $p < 0.5$. If F and G are lognormal, then p will get even smaller, since the lognormal is even more skewed. Now consider the BF model, with $\rho > 1$ and F and G both lognormal with a common median zero. Then instead of centering (U/mn) at the much smaller value p, FP centers it at the much larger value 0.5. Clearly, the limiting distribution will have the wrong mean.

This was also found in earlier simulation studies. Here is an example. Let F and G be exponential with a common median zero and scale parameter 1 and 2 respectively. For a two-sided test with nominal level 10 percent, and $n = m = 20$, the empirical levels were about 7.5 percent and 2.5 percent respectively. Increasing the sample size to $n = m = 30$, only made matters worse, the levels becoming about 9 percent and 1 percent respectively.

The Mann-Whitney statistic U also has this drawback. Its mean under the location model is $(0.5)mn$. But in the BF model, the correct mean is pmn, involving the unknown probability p. Therefore, the test now uses the wrong mean (besides the wrong variance, as explained earlier) and thus becomes highly non-robust.

The BP procedure achieves robustness to skewness by proposing a new estimate of p, say $\hat{p}$, and estimates the quantiles of $\sqrt{m+n}(\hat{p} - p)$ by bootstrapping. However, the consistency of the bootstrap estimate requires F and G to have the same shape, thereby rendering it inapplicable to the GBF problem. Moreover, due to the somewhat messy nature of this estimate $\hat{p}$, convergence to the limiting distribution slows down in the case of extremely unbalanced samples involving negative pairing; that is, the larger scale parameter occurring with the smaller sample size. For example, for a two-sided test at nominal level 5% and sample sizes $n = 10$ and $m = 20$ and the standard deviation of F being twice as much as that of G, the empirical levels for the normal, exponential and lognormal distributions were about 10%, 12% and 15% respectively.

5.2. Monte Carlo results for the two-sample setting. In view of the above discussion, only our test and the Mann-Whitney test were studied. The latter, despite its non-robustness, was included because of its computational simplicity.

Samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ were drawn from distributions F and G (described a little later), where $(n, m) = (10,10)$, $(10,20)$ and $(20,20)$. Two-sided tests with nominal level 5% were performed. The level was estimated by Monte Carlo rejection rates based on $50,000$ simulation runs. Each bootstrap was based on $2,5000$ repetitions. The power was estimated in the same vain, where the alternative was specified by a location shift of size 0.5, i.e. that constant was added to the observations of the X-sample.

The following distributions for F and G were taken into account.

i) Short-tailed distribution: We studied a short-tailed distribution, namely a mixture of $U(-0.5, 0.5)$, the uniform distribution on $(-0.5, 0.5)$, and a normal distribution, $N(a, b)$, with specific choices of the mean a and the standard deviation b. Indeed, (Cox 1977, p. 1083), when analyzing textile data, encountered short-tailed sets as frequently as long-tailed ones for that problem. Hogg (1974, p. 976) and Wegman and Carrol (1977, p. 976) reported similar experiences in connection with data on examination scores and data passing through electronics instruments in the US Naval Laboratory. Both short-tailed and long-tailed sets arise in the Physics and Astronomy sets of Stigler (1977) and the Analytical Chemistry sets sent to us by Rocke, Downs and Rocke (1982). Moreover a mixture of $U(-0.5, 0.5)$ and $N(0, 1)$ is a realistic short-tailed distribution (cf. Wegman and Carrol 1977, p. 976). Therefore, we chose $(0.5)U(-0.5, 0.5)+(0.5)N(0, 1)$ as a typical short-tailed distribution.

ii) Normal Mixtures: Normal mixtures provide a common approach to mimic long-tailed distributions. We considered the mixture model $pN(0, 1)+(1-p)N(0, 4)$ for $p \in \{1, 0.9, 0.85\}$.

iii) Asymmetric Distributions: The standard exponential and the standard log-normal are examples of asymmetric distributions. These were chosen due to their importance in biostatistics, industrial engineering and reliability studies. In the case of symmetric distributions, we may assume that, without loss of generality, the pseudo-medians are zero. Clearly, this is no longer true for skewed distributions. Now we have to find the pseudo-median δ and add it to observations of the Y-sample. Thus, to ensure H_0, we have to work with the samples $X_1, \ldots, X_n$ and $Y_1 + \delta, \ldots, Y_m + \delta$. According to Hamza (2008), this can be achieved by solving a non-linear algebraic equation, if F and G are exponential, but not for log-normal distributions. Thus, δ was obtained by simulation.

iv) Generalized extreme value distribution: We also investigated the performance of our proposal for some members of the generalized extreme value distribution

Table 1: *Simulated level under various distributions for equal scales as well as under heteroscedasticity for the Wilcoxon test (W) and the proposed test (PS).*

		$n = m = 10$		$n = m = 20$		$n = 10, m = 20$	
Distribution	*Model*	*W*	*PS*	*W*	*PS*	*W*	*PS*
St. Normal	1	4.4	4.9	5.3	5.2	4.8	5.5
	2	4.1	4.8	5.2	5.4	6.0	5.4
	3	4.4	4.5	6.2	5.0	8.4	5.4
Mixture	1	3.1	4.5	5.4	6.3	4.5	5.7
	2	3.4	4.4	4.7	5.5	6.4	5.6
	3	4.5	4.5	5.6	5.8	7.6	6.1
10% Cont.	1	4.0	5.3	4.4	4.6	4.6	6.0
	2	4.5	5.5	4.3	4.6	6.0	5.0
	3	4.4	4.6	4.9	4.6	7.4	4.9
15% Cont.	1	4.3	5.6	5.6	5.4	5.5	5.7
	2	4.5	5.7	5.7	5.3	6.5	6.4
	3	4.7	4.8	6.1	5.6	8.8	6.2
Exponential	1	3.8	4.4	4.9	4.9	5.3	6.5
	2	5.0	5.1	6.6	4.7	8.7	7.3
	3	6.4	5.0	9.6	5.1	11.5	8.1
Lognormal	1	3.8	4.9	5.4	4.9	7.5	7.0
	2	4.1	4.3	7.1	5.1	9.3	8.8
	3	5.5	5.5	7.7	5.0	12.6	8.4
Nor-Lognor	1	6.1	5.7	5.9	4.1	5.5	5.1

(GEV), i.e. for the family of distributions given by the d.f.s

$$F_{GEV}(x) = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right), \qquad x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ is the location, $\sigma > 0$ the scale and $\gamma \in \mathbb{R}$ the shape parameter. As well known, the GEV family covers the Weibull ($\gamma = 0$) and Fréchét ($\gamma > 0$) distributions as special cases. We considered the GEV distributions corresponding to $\xi = 0,\ 0.25,\ , 0.5,\ 0.75$. The balanced case was studied for $n = m \in \{15, 25\}$, as an unbalanced design we selected $n = 25$ and $m = 50$. Again,we studied models 1 to 3 corresponding to homogeneous variances and two heteroscedastic settings.

Whereas Table 1 and Table 2 summarize our findings for the distributions i) - iii), the results for the GEV distributions are provided in Tables 3 and 4. For each distribution F, we considered three scale models. Model 1 corresponds to equal scales, whereas in model 2 and model 3 the ratio of the standard deviation of the second sample to the first sample equals $\sqrt{2}$ and 2, respectively.

Table 2: *Simulated power of the Wilcoxon test (W) and the proposed test (PS) for selected distributions corresponding to the Behrens-Fisher and Generalized Behrens-Fisher models.*

		$n = m = 10$		$n = m = 20$		$n = 10, m = 20$	
Distribution	*Model*	W	PS	W	PS	W	PS
St. Normal	1	18.4	20.0	31.4	32.2	24.6	26.1
	2	14.7	16.4	23.4	22.7	19.8	17.7
	3	9.0	9.7	17.5	16.2	16.2	12.5
Mixture	1	36.3	34.8	68.7	60.0	50.8	46.2
	2	26.1	24.3	54.2	45.9	39.6	32.5
	3	18.2	17.6	38.1	32.5	29.1	21.3
10% Cont.	1	12.2	12.1	27.7	25.1	15.2	16.2
	2	9.0	9.5	21.0	19.7	13.1	11.6
	3	8.0	9.2	14.2	14.2	11.3	8.5
15% Cont.	1	12.4	13.6	24.7	23.0	14.9	13.9
	2	9.7	11.6	19.7	19.2	12.8	11.3
	3	9.3	9.4	12.2	12.0	11.9	9.0
Exponential	1	27.8	24.4	58.5	44.3	49.9	35.3
	2	17.3	17.2	39.3	34.8	32.5	25.7
	3	7.9	10.5	13.6	22.0	16.7	16.0
Lognormal	1	19.1	14.2	37.5	24.1	32.4	20.7
	2	10.9	10.6	19.4	17.4	21.2	15.2
	3	5.1	8.0	10.8	16.2	11.5	11.0
Nor-Lognor	1	21.3	14.5	36.7	22.3	25.2	14.4

Table 3: *Simulated level of the Wilcoxon test (W) and the proposed test (PS) for some selected GEV distributions.*

		$n = m = 15$		$n = m = 25$		$n = 25, m = 50$	
Distribution	*Model*	W	PS	W	PS	W	PS
GEV $\gamma = 0$	1	5.1	4.9	5.0	5.3	5.0	5.4
	2	5.3	4.7	5.6	5.3	4.1	5.1
	3	6.3	5.1	6.7	5.2	3.9	5.1
GEV $\gamma = 0.25$	1	5.0	4.6	5.1	4.8	5.0	5.2
	2	5.8	4.7	5.8	4.7	4.6	4.9
	3	7.2	4.9	7.8	5.1	5.7	4.6
GEV $\gamma = 0.5$	1	5.0	4.2	5.0	4.7	4.9	5.2
	2	5.8	4.4	6.0	4.6	5.0	4.6
	3	7.7	5.5	9.4	4.9	6.6	4.6
GEV $\gamma = 0.75$	1	5.1	4.0	5.0	4.8	5.1	5.1
	2	6.4	4.1	6.5	4.3	5.4	4.6
	3	9.1	5.5	11.0	4.6	8.8	4.3

Table 4: *Simulated power of the Wilcoxon test (W) and the proposed test (PS) for selected GEV distributions.*

		$n = m = 15$		$n = m = 25$		$n = 25, m = 50$	
Distribution	*Model*	W	PS	W	PS	W	PS
GEV $\gamma = 0$	1	61.7	58.6	82.0	79.4	91.6	88.8
	2	42.5	43.7	63.5	65.9	77.5	81.9
	3	25.1	28.3	39.2	46.0	49.3	67.4
GEV $\gamma = 0.25$	1	54.5	43.9	76.9	65.2	86.6	75.4
	2	37.6	34.2	57.5	53.9	70.3	69.7
	3	19.0	21.3	29.5	35.2	36.0	54.6
GEV $\gamma = 0.5$	1	49.0	32.6	71.0	50.3	81.2	61.3
	2	33.1	25.5	53.2	43.0	64.1	57.4
	3	15.0	16.0	23.0	26.9	28.9	44.7
GEV $\gamma = 0.75$	1	45.4	25.4	66.1	38.7	77.3	51.0
	2	31.3	20.1	48.4	32.7	59.6	47.3
	3	13.5	13.1	18.8	20.7	19.4	33.7

5.3. Monte Carlo results for many-to-one comparisons. We also investigated the behavior of the bootstrap test for the many-to-one ANOVA testing problem for small sample sizes. However, in view of the unsatisfactory performance of the Wilcoxon test in the two-sample case (as explained later), we focused on level and power of our test. We considered three samples, i.e., in the notation of Section 2, the null hypothesis is given by the set $\mathcal{H}_0 = \{(1,2), \ldots, (1,a)\}$ with $a = 3$. Independent random samples

$$X_{1i} \sim F(x/\sigma_1), \qquad X_{2j} \sim F((x - \delta_2 - \Delta_2)/\sigma_2), \qquad X_{3k} \sim F(x - \delta_3 - \Delta_3),$$

of sample sizes n_1, n_2, n_3 were simulated, where for $i = 1, 2$ δ_i denotes the pseudo-median of the X_i-sample relative to the X_1-sample (control group) for fixed σ_1, σ_2 and $\Delta_2 = \Delta_3 = 0$, where Δ_2 and Δ_3 are the location shifts. This means, $\Delta_2 = \Delta_3 = 0$ corresponds to the null hypothesis of identical pseudo-medians. For Model 1, $\sigma_1 = \sigma_2 = 1$ (equal scales), whereas model 2 corresponds to $\sigma_1 = 2$ and $\sigma_2 = \sqrt{2}$. The power was assessed using $\Delta_2 = \Delta_3 = 1$.

Table 5 provides the results for the small sample size setting, i.e., $n_1 = n_2 = n_3 = 10$. Each entry is based on 10,000 simulation runs and a bootstrap with 1,000 replications.

5.4. Discussion of results. The simulation results support the following conclusions.

Behrens-Fisher model:

We start our discussion with the Wilcoxon test.

Table 5: *Empirical size and power of the many-to-one ANOVA test.*

Distribution	*Model*	*Level*	*Power*
Mixture	1	4.71	91.2
	2	4.97	53.3
Exponential	1	4.55	72.5
	2	6.03	35.4
10% Cont.	1	4.60	49.96
	2	4.61	16.24
Lognormal	1	3.74	49.24
	2	5.31	24.46

(i) Symmetric distributions. For $n = m = 10$, the Wilcoxon test was quite robust even for the Behrens-Fisher models 2 and 3. This observation is not surprising, since due to the smallness of the samples, the differences between the critical levels for the Behrens-Fisher and location models were not significant. However, they became significant with increasing sample sizes, as shown by the liberal levels for $n = m = 20$.

(ii) Skewed distributions. Multiplying a standard exponential (or lognormal) variate by a positive constant to effect a scale change, automatically induced a change in location as well. As a result, the test became liberal, unless the samples were extremely small. However, the actual performance was even worse. The test became extremely liberal for lower tails and conservative for the upper tails, so that some kind of an averaging effect took place for a two-tailed test. It is enough to explain this for the case of lognormal (model 3) with $n = m = 20$. For one-sided testing at nominal level 0.05, the empirical levels were found to be 0.15 and 0.02 for the lower- and upper-tailed tests respectively.

Our test performed creditably both when the samples were small ($n = m = 10$) and when they were moderate ($n = m = 20$). One plausible explanation for the slightly worse performance when $n = 10$ and $m = 20$ is that the extreme imbalance in the relatively small samples, coupled with negative pairing, somewhat slowed down convergence to the asymptotic results.

Generalized Behrens-Fisher model (Normal-Lognormal):

Now the Wilcoxon test seemed to perform reasonably well. However, this was a case of two wrongs making one right. Further studies showed that it was liberal for the upper tail and conservative for the lower tail. For example, when $n = m = 20$ and nominal level was 0.05, the empirical levels for the lower- and upper-tailed tests were given by 0.02 and 0.095 respectively. Our test performed creditably for all sample sizes.

Generalized extreme value distribution:

For distributions of the GEV type or similar to them, our test makes really a strong point. The empirical sizes in Table 3 clearly demonstrate that the Mann-Whitney test is no longer applicable when it comes to GEV distributions and heteroscedasticity. For a scale factor of 2 (model 3) the empirical level is far from being acceptable. As a consequence, its power is better in some cases as can be seen from Table 4. However, in the heteroscedastic case even power drops severely and our proposal becomes more powerful. Thus, for GEV-like distributions the methodology developed in the present substantially outperforms the classical approach.

Many-to-one comparison:

In view of the unsatisfactory performance of the Wilcoxon test, only the many-to-one comparison version of our test was considered. As the simulation results show, the behavior of our test for the (generalized) Behrens-Fisher problem provides a very reliable and powerful statistical test.

6 Examples

We applied our test procedure to some real data sets from engineering, photovoltaics, physics, chemistry and psychology. Firstly, to illustrate our procedure and its applicability to a wide range of scientific areas. Secondly, to gain further insight into these interesting data as some of them already appeared in the literature. Last but not least, since those data sets, and some others we cannot publish, motivated our work on that classic statistical problem. For our analyses, which were conducted in R, the bootstrap critical values and bootstrap p-values were estimated by the Monte Carlo method using 500,000 independent replications. The random number generator was initialized with the seed 1. For testing the equality of scales the F-K:med test was used (see Fligner and Killen, 1976 and Hall and Padmanabhan, 1997).

6.1. Electrical Engineering. Nair (1984) gives the following two-sample data set from electrical engineering. It consists of the times (in minutes) to breakdowns of an insulating fluid under elevated voltage stresses of 32 kV (X-sample),

$$0.27, 0.40, 0.69, 0.79, 2.75, 3.91, 9.88, 13.95, 15.93,$$
$$27.30, 53.24, 82.85, 89.25, 100.58, 215.50,$$

and under 36 kV (Y-sample),

$$0.35, 0.59, 0.96, 0.99, 1.69, 1.97, 2.07, 2.58, 2.71,$$
$$2.90, 3.67, 3.99, 5.35, 13.77, 25.50.$$

The analysis in Hall and Padmanabhan (1997) revealed a scale difference. The value of our test statistic is 1.824234. The bootstrap 5% lower and upper critical value are -1.440128 and 1.301188, respectively, and the 5% two-sided bootstrap critical

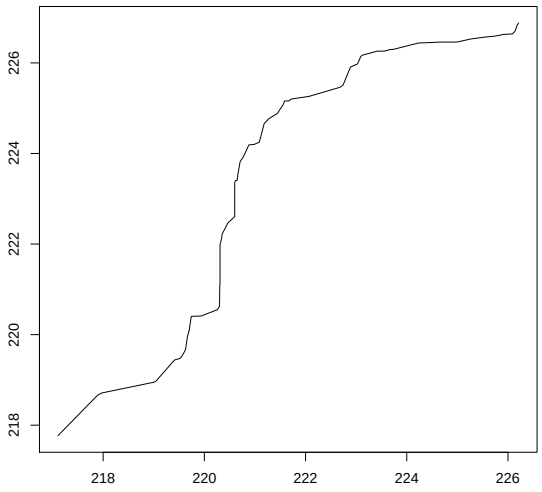

Figure 1: QQ-plot of photovoltaic measurements of two shipments.

value is 1.949623. The bootstrap p-value for the one-sided test $H_0 : \delta \leq 0$ against $H_1 : \delta > 0$ is 0.0075. Hence, we may conclude that the pseudo-median is positive indicating larger values in the first sample. The pseudo-median is estimated by $\widehat{\delta}_N = 26.345$.

6.2. Photovoltaics. In photovoltaics the power output of photovoltaic modules is the most important variable to assess quality, see Steland and Zähle (2009) and Herrmann and Steland (2010). The physical production process and further technical issues imply that one cannot specify a parametric class of distributions. Indeed, almost any type of distribution (heavily skewed, symmetric, bimodal etc.) is observed in practice. Further, the distribution of photovoltaic modules taken from different lots or shipments may differ, even if these modules satisfy the same technical specifications. Figure 1 depicts an empirical QQ-plot of two random samples of sample sizes $n = 30$ and $m = 50$ drawn from two shipments of the same module for a solar power plant. Here the main issue is that the distributions are quite different. The pseudo-median provides a convincing approach to analyze these samples. We applied our approach to check whether the samples differ in location. The value of our test statistic is -2.389 and the pseudo-median is estimated by $\widehat{\delta}_N = -1.99$. That difference is significant at any reasonable level, since the bootstrapped p value is 0.0022.

6.3. Physics. The following classic data sets taken from Cressie (1997, p. 46) pertain to the Heyl and Cook (1936) measurements of the acceleration of gravity, expressed as deviations from $980,060 \times 10^3 cm/sec^2$. Heyl and Cook described the great amount of care taken in the experiments and the adjustments to avoid systematic error. Three of the eight series are as follows:

$$\begin{aligned} x_1 &= (87, 95, 98, 100, 109, 100, 81, 75, 68, 67), \\ x_2 &= (78, 78, 78, 86, 87, 81, 73, 67, 75, 82, 83), \\ x_3 &= (84, 86, 85, 82, 77, 76, 80, 83, 81, 78, 78, 78). \end{aligned}$$

These series of measurements are known to exhibit inhomogeneity in their variances. The values of the F-K:med statistic and its 95% quantiles were 168 and 103 respectively, resulting in rejection of the hypothesis of equal scales, thus suggesting a Behrens-Fisher model. The value of our statistic and its bootstrap 95% quantiles are 2.999986 and 3.0328 respectively, confirming that there is no difference in location.

6.4. Psychology. The following data sets based on the skin resistance of three groups, are a slight modification of the original data sets provided by R. Wilcox. Those observations were collected in a study dealing with schizophrenia. The three samples are

$$\begin{aligned} x_1 &= (0.998, 0.469, 0.53, 0.558, 0, 0, 0, 0, 0.282, 2.680), \\ x_2 &= (0.250, 0, 0, 0.390, 0.348, 0, 0.207, 0.444, 0, 0.318), \\ x_3 &= (0.250, 0, 0, 0, 0, 0.115, 0.795, 0.177, 0, 0.158, 0). \end{aligned}$$

The values of the F-K:med statistic and its 95% quantiles were 116 and 94 respectively, resulting in rejection of the hypothesis of equal scales. So, once again Behrens-Fisher model seems appropriate.

Next, the value of our statistic and its bootstrap 95% quantiles are 2.116502 and 3.025593 respectively. Hence, the null hypothesis is accepted.

6.5. Analytical Chemistry. Finally, we re-analyzed one of the data set of chemical measurements collected and analyzed in Rocke, Downs and Rocke (1982). The data consist of two independent samples with sample sizes 24 and 25, respectively. They have been analyzed and provided in Hill and Padmanabhan (1991), thus we do not reproduce them here. The data exhibit asymmetry, lighter tails than a normal distribution as well as heteroscedasticity. Applying our methodology leads to a value for the test statistic of -2.893042. The bootstrap critical value is 1.754018 and the bootstrap p value $3.8 \cdot 10^{-5}$ leading to a clear rejection of the null hypothesis. The pseudo-median is estimated by -0.13 which is in close agreement to the difference of the arithmetic means -0.1353667.

Acknowledgement. We are grateful to Professor Rob Hyndman of the Department of Econometrics and Business Statistics, Monash University, for his interest

and encouragement, Dr. Kais Hamza of the Department of Mathematics, Monash University, for some useful conversations, Dr. Werner Herrmann, TÜV Rheinland Group, Cologne, Germany, for the photovoltaic data, and Professor Wilcox of the Department of Psychology, University of Southern California, for the psychology data, and Professor David Rocke, University of California at Davis, for the data set on Analytical Chemistry. We also thank Dipl.-Math. Sabine Weidauer for proofreading an early draft.

References

Abdus, B., Ghoudi, K. and Remillard, B. (2003). Nonparametric weighted symmetry tests. *Canad. J. Statist.*, **31**, 357–381.

Babu, G.J. and Padmanabhan, A.R. (2002). Re-sampling methods the non-parametric Behrens-Fisher problem. *Sankhyā A*, **64**, 678–692.

Babu, G.J. and Padmanabhan, A.R. (2007). Re-sampling methods for testing for location against unrestricted and ordered alternatives. *J. Statist. Plann. Inference*, **137**, 3261–3267.

Babu, G.J., Padmanabhan, A.R. and Puri, M.L. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biom. J.*, pp. 321–339.

Babu, G.J. and Rao, C.R. (1993). Bootstrap Methodology. In *Computational statistics*, (C.R. Rao, ed.), 627-659. Handbook of Statistics, **9**. North-Holland, Amsterdam.

Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196–1217.

Bickel, P.J. and Lehmann, E.L. (1975). Descriptive statistics for nonparametric models ii. location. *Ann. Statist.*, **3**, 1045–1069.

Boos, D.D. and Brownie, C. (1991). Mixture models for continuous data in dose-response studies when some animals are affected by treatments. *Biometrics*, **47**, 1489–1504.

Boos, D.D. and Brownie, C. (2004). Comparing variances and other measures of dispersion. *Statist. Sci.*, **19**, 571–578.

Borovskikh, Y.V. (1996). *U-statistics in Banach spaces*. VSP, Utrecht.

Chow, S.C. and Liu, J.P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, New York.

Cox, D.R. (1977). Discussions on 'Do robust estimators work with real data?' by S.M. Stigler. *Ann. Statist.*, **5**, 1055–1098.

Cressie, N. (1997). Jackknifing in the presence of inhomogeneity. *Technometrics*, **39**, 45–51.

Dehling, H., Denker, M. and Woyczynski, W.A. (1990). Resampling U-statistics using p-stable laws. *J. Multivariate Anal.*, **34**, 1–13.

Denker, M. (1985). *Asymptotic distribution theory in nonparametric statistics*. Advanced Lectures in Mathematics. Friedr. Vieweg & Sohn, Braunschweig.

Fligner, M.A. and Killen, T.J. (1976). Distribution free two sample test for scale. *J. Amer. Statist. Assoc.*, **71**, 210–213.

Fligner, M.A. and Policello, G.E. (1981). Robust rank procedures for the Behrens-Fisher problem. *J. Amer. Statist. Assoc.*, **76**, 162–167.

Gill, R. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (part1). *Scand. J. Stat.*, **16**, 97–128.

Hall, P. and Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.*, **37**, 3929–3959.

Hall, P. and Padmanabhan, A.R. (1997). Adaptive inference for the two-sample scale problem. *Technometrics*, **39**, 412–422.

Hamza, K. (2008). Personal communication.

Herrmann, W. and Steland, A. (2010). Evaluation of photovoltaic modules based on sampling inspection using smoothed empirical quantiles. *Progress in Photovoltaics*, **18**, 1–9.

Hettmansperger, T.P. (1991). *Statistical Inference Based on Ranks*. John Wiley & Sons, Inc., New York.

Heyl, P. and Cook, G. (1936). The value of gravity in Washington. *J. Res. U.S. Bureau Stand.*, **17**, 805–839.

Hill, N.J. and Padmanabhan, A.R. (1991). Some adaptive robust estimators which work with real data. *Biom. J.*, **33**, 81–101.

Hodges, J.L. and Lehmann, E.L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.*, **34**, 598–611.

Hogg, R.V. (1974). Adaptive robust procedures. *J. Amer. Statist. Assoc.*, **69**, 909–927.

Hušková, M. and Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *Ann. Statist.*, **21**, 1811–1823.

Lamb, R.H., Boos, D. D. and Brownie, C. (1996). Testing for effects on variance in experiments with factorial treatment structure and nested errors. *Technometrics*, **38**, 170–177.

Lee, A.J. (1990). *U-statistics*. Statistics: Textbooks and Monographs, Vol. 110, Marcel Dekker Inc., New York.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden Day.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychol. Bull.*, **105**, 156–166.

Nair, V. (1984). On the behavior of some estimators from probability plots. *J. Amer. Statist. Assoc.*, **17**, 823–830.

Podgor, M.J. and Gastwirth, J. (1964). On nonparametric and generalized tests for the two-sample problem with location and scale alternatives. *Stat. Med.*, **13**, 747–758.

Puri, M.L. and Sen, P.K. (1971). *Nonparametric Methods in Multivariate Statistical Analysis*. John Wiley and Sons, New York.

Randles, R.H. and Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. John Wiley and Sons, New York.

Rocke, D.M., Downs, G.W. and Rocke, A.J. (1982). Are robust estimators really necessary? *Technometrics*, **24**, 95–101.

Rust, S. and Fligner, M.A. (1984). A modification of the Kruskal Wallis test for the generalized Behrens-Fisher problem. *Comm. Statist. Theory Methods*, **13**, 2013–2027.

Shao, J. and Yu, D. (1995). *The Jacknife and Bootstrap*. Springer Series in Statistics. Springer, Heidelberg.

Shorack, G.R. (2000). *Probability for statisticians*. Springer Texts in Statistics. Springer-Verlag, New York.

Steland, A. (1997). On a rank test in a two-factor model with varying dependent repeated measurements. *J. Nonparametr. Stat.*, **8**, 215–235.

Steland, A. (1998). Bootstrapping rank statistics. *Metrika*, **47**, 251–264.

Steland, A. (2005). On the distribution of the clipping median under a mixture model. *Statist. Probab. Lett.*, **71**(1), 1–13.

Steland, A. and Zähle, H. (2009). Sampling inspection by variables: nonparametric setting. *Stat. Neerl.*, **63**(1), 101–123.

Stigler, S.M. (1977). Do robust estimators work with real data? *Ann. Statist.*, **5**, 1055–1098.

van der Vaart, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 3. Cambridge University Press, Cambridge.

WEGMAN, E.J. and CARROLL, R.J. (1977). A Monte Carlo study of robust estimators of location. *Comm. Statist. Theory Methods*, **A6**, 795–812.

WILCOX, R.R. (1987). New designs in analysis of variance. *Ann. Rev. Psychol.*, **49**, 163–170.

WILCOX, R.R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means and designing simulation studies. *British J. Math. Statist. Psych.*, **49**, 163–170.

ZUMBO, B.D. (2002). An adaptive inference strategy: The case of auditory data. *J. Mod. Appl. Statist. Methods*, **1**, 61–68.

ZUMBO, B.D. and KOH, K.H. (2005). Manifestation of differences in item-level characterestics of scale measurements, involving tests of multi-group confirmatory factor analyses. *J. Mod. Appl. Statist. Methods*, **4**, 275–282.

A Proofs

Although at first glance the proposed bootstrap scheme seems to be rather straightforward, it is non-standard in that we resample from differences which are, firstly, dependent with a certain structure, and, secondly, asymmetrically adjusted for a location difference by a nonlinear robust statistic of the data, which complicates the probabilistic analysis. The latter means that our bootstrap resamples from a specific set of residuals.

Therefore, this appendix establishes the required theoretical results for the proposed methods and provides some additional related results such as laws of large numbers and approximation results which are of independent interest. To the best of our knowledge, both the bootstrap results for such a setting and the central limit theorem for the signed rank statistic applied to the between-samples differences are new. However, it turns out that the counting structure of the statistics allows us to base our proofs on the theory of U-statistics. We shall introduce notation and cite required results when needed; for general background on required probabilistic results on U-, V- and rank statistics, we refer to Borovskikh (1996), Denker (1985), Lee (1990), Randles and Wolfe (1979) and Shorack (2000).

A.1. Proof of the bootstrap central limit theorems for the two-sample setting. To establish bootstrap central limit theorems we use the fact that a signed rank statistic is asymptotically equivalent to an U-statistic. Bootstrap central limit theorems have been extensively studied in the literature, we refer to Bickel and Freedman (1981), Dehling, Denker and Woyczynski (1990), and Hušková and Janssen (1993), among many others. For a different approach to show consistency of the bootstrap see Steland (1998). However, as discussed in Subsection 3.3., those results can not be directly applied to our problem.

To fit the one-sample framework, define the dependent random variables

$$Z_{(i-1)m+j} = X_i - Y_j, \qquad i = 1, \ldots, n,\ j = 1, \ldots, m.$$

Recall that the sum of the signed ranks,

$$W_N' = \sum_i R_i \mathbf{1}(Z_i > 0)$$

is not an U-statistic, but can be written as a linear combination of U-statistics, since

$$W'_N = \sum_i \mathbf{1}(Z_i > 0) + \sum_{i<j} \mathbf{1}(Z_i + Z_j > 0).$$

The first term corresponds to the kernel $h_1(z) = \mathbf{1}(z > 0)$ and the associated U statistic $\overline{U}_N = N^{-1} \sum_{i=1}^N \mathbf{1}(Z_i > 0)$, whereas the second term is related to the U statistic

$$\widetilde{U}_N = \binom{N}{2}^{-1} \sum_{i<j} \mathbf{1}(Z_i + Z_j > 0)$$

given by the kernel $h_2(x, y) = \mathbf{1}(x + y > 0)$, $x, y \in \mathbb{R}$. Hence,

$$\begin{aligned} T_N &= \sqrt{n+m} \binom{N}{2}^{-1} \left(W'_N - \frac{N(N+1)}{4} \right) \\ &= \sqrt{n+m} \left(\widetilde{U}_N + \binom{N}{2}^{-1} N \overline{U}_N - 1/2 \right) \\ &= \sqrt{n+m} (\widetilde{U}_N - 1/2) + o_P(1). \end{aligned} \tag{A.1}$$

To verify a central limit theorem for the one-sample U-statistic $\widetilde{U}_N$ based on dependent observations, we will approximate it by the two-sample U-statistic

$$U_N = \binom{n}{2}^{-1} \binom{m}{2}^{-1} \sum_{1 \le i < i' \le n,\ 1 \le j < j' \le m} \mathbf{1}(X_i - Y_j + X_{i'} - Y_{j'} > 0).$$

U_N is induced by the two-sample kernel

$$h(x_1, x_2, y_1, y_2) = \mathbf{1}(x_1 - y_1 + x_2 - y_2 > 0), \tag{A.2}$$

for $x_1, x_2, y_1, y_2 \in \mathbb{R}$, which is symmetric in its $x-$ and $y-$ arguments, respectively. Obviously, U_N is a non-degenerate U-statistic and estimates the parameter

$$\theta = \theta(F, G) \quad = \quad EU_N = P(X_1 - Y_1 + X_2 - Y_2 > 0).$$

Consider now the one-sample U-statistic defining the dominant term of the statistic T_N for the sample $D_{ij} = X_i - Y_j$,

$$\widetilde{U}_N = \binom{N}{2}^{-1} \sum_{(i-1)m+j < (i'-1)m+j'} \mathbf{1}(X_i - Y_j + X_{i'} - Y_{j'} > 0).$$

Here the sum extends over all $i, i' = 1, \ldots, n$ and $j, j' = 1, \ldots, m$ such that $(i-1)m + j < (i'-1)m + j'$. The following lemma states that $\widetilde{U}_N$ is asymptotically equivalent to U_N, and holds true for the original as well as the bootstrap version of the statistic.

LEMMA A.1. *Assume $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent i.i.d. samples defined on a common probability space $(\Omega, \mathcal{A}, \mu)$ with distributions F and G, respectively, under the probability measure μ. Then, under μ*

$$\sqrt{n+m}(\widetilde{U}_N - \theta(F,G)) = \sqrt{n+m}(U_N - \theta(F,G)) + o(1), \qquad \mu - a.s.,$$

as $n+m \to \infty$ such that $n/(n+m) \to \lambda \in (0,1)$.

PROOF. Note that $(i-1)m + j < (i'-1)m + j'$ if and only if either $i < i'$ and j, j' arbitrary or $i = i'$ and $j < j'$. By symmetry of the kernel h associated with U_N,

$$\begin{aligned}
\binom{N}{2}\widetilde{U}_N &= \sum_{(i-1)m+j<(i'-1)m+j'} \mathbf{1}(X_i - Y_j + X_{i'} - Y_{j'} > 0) \\
&= 2 \sum_{1\le i<i'\le n, 1\le j<j'\le m} \mathbf{1}(X_i - Y_j + X_{i'} - Y_{j'} > 0) \\
&\quad + \sum_{1\le i<i'\le n, 1\le j\le m} \mathbf{1}(X_i - Y_j + X_{i'} - Y_j > 0) \\
&\quad + \sum_{1\le i\le n, 1\le j<j'\le m} \mathbf{1}(X_i - Y_j + X_i - Y_{j'} > 0).
\end{aligned}$$

The first term equals $2\binom{n}{2}\binom{m}{2}U_N$. The second term has $n(n-1)m/2$ summands and the third one sums up $nm(m-1)/2$ bounded terms. Hence, noting that $\frac{\binom{n}{2}\binom{m}{2}}{\binom{N}{2}} = \frac{1}{2} - \frac{n+m-2}{2(nm-1)}$ and, since the summands are bounded by 1,

$$\sqrt{n+m}\frac{n+m-2}{2(nm-1)}U_N = o(1), \qquad \mu - a.s.,$$

(even for all $\omega \in \Omega$), if $n/(n+m) \to \lambda \in (0,1)$, we obtain

$$\begin{aligned}
\sqrt{n+m}(\widetilde{U}_N - \theta(F,G)) &= \sqrt{n+m}\left(2\frac{\binom{n}{2}\binom{m}{2}}{\binom{N}{2}}U_N - \theta\right) \\
&\quad + O\left(\frac{n(n-1)m}{(nm)(nm-1)}\right) + O\left(\frac{m(m-1)n}{(nm)(nm-1)}\right) \\
&= \sqrt{n+m}(U_N - \theta(F,G)) + o(1),
\end{aligned}$$

as $n+m \to \infty$ such that $n/(n+m) \to \lambda \in (0,1)$, μ-a.s. □

REMARK A.1. In Lemma A.1, μ may be a random probability measure such as a (regular) conditional distribution. Indeed, we shall apply Lemma A.1 using $\mu = P$ as well as with $\mu = P^*$.

It is worth mentioning that a consequence of Lemma A.1 and (A.1) is that

$$(2/N^2)W'_N = \widetilde{U}_N + o_P(1) = \theta(F,G) + o_P(1), \qquad n, m \to \infty,$$

which proves the following result.

PROPOSITION A.1. *A weakly consistent estimator of $\theta(F,G)$ is given by*

$$\widehat{\theta}(F,G) = (2/N^2)W'_N.$$

Next let us first verify the strong consistency of the estimator $\widehat{\delta}_N$. Define the d.f.

$$K(x) = P((X_1 - Y_1 + X_2 - Y_2)/2 \le x), \qquad x \in \mathbb{R}.$$

Assume $K^{-1}(x)$ is continuous at $1/2$. We will now verify the assertion of Theorem 3.2, namely that for $n, m \to \infty$,

$$\widehat{\delta}_N \overset{P-a.s.}{\to} \delta = K^{-1}(1/2).$$

PROOF OF THEOREM 3.2. Note that $\widehat{\delta}_N = K^{-1}_{n,m}(1/2)$, where

$$K_{n,m}(x) = \frac{1}{n^2m^2} \sum_{i,k=1}^{n} \sum_{j,l=1}^{m} \mathbf{1}((X_i - Y_j + X_k - Y_l)/2 \le x).$$

$K_{n,m}(x)$ is a two-sample V-statistic based on the i.i.d. samples $\{X_i/2\}$ and $\{Y_j/2\}$. Thus, for each fixed x $K_{n,m}(x) \to E\mathbf{1}((X_1 - Y_1 + X_2 - Y_2)/2 \le x) = K(x)$, as $n, m \to \infty$. Using arguments as in the proof of Theorem 3 of Steland (2005) one can also show uniform convergence, but we shall not elaborate on this issue. Since $K_{n,m}(x) \to K(x)$ in all continuity points x of K is equivalent to $K^{-1}_{n,m}(t) \to K^{-1}(t)$ for all t where K^{-1} is continuous (e.g. van der Vaart, 1998, 21.2), the assertion follows. □

Let us now verify asymptotic normality of T_N.

PROOF OF THEOREM 3.1. Using the decomposition (A.1) as well as Lemma A.1 with $\mu = P$ we obtain that

$$\sqrt{n+m}\left\{ \binom{N}{2}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij}\mathbf{1}(D_{ij} > 0) - \theta(F,G) \right\}$$

equals

$$\sqrt{n+m}(U_N - \theta(F,G)) + o(1) + o_P(1), \qquad \mu - a.s.$$

U_N is a two-sample U-statistic calculated from the two independent i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. Hence, we may apply the central limit theorem for two-sample U-statistics (e.g. Randles and Wolfe, 1979, Theorem 3.4.14) to conclude that under fixed distributions F and G

$$\sqrt{n+m}(U_N - \theta(F,G)) \overset{d}{\to} N(0, \eta^2(F,G)),$$

as $n, m \to \infty$ with $n/(n+m) \to \lambda \in (0,1)$. Notice that

$$\sigma_{01}(F,G) = \mathrm{Cov}\left(h(X_1, X_2, Y_1, Y_2), h(X_1, X_3, Y_3, Y_4)\right)$$

and

$$\sigma_{10}(F,G) = \mathrm{Cov}\,(h(X_1,X_2,Y_1,Y_2), h(X_3,X_4,Y_1,Y_3)).$$

where h is given by (A.2). □

Let us now turn to the verification of the proposed bootstrap procedure. In the sequel, we allow for general bootstrap sample sizes n^* and m^*, respectively. Also define

$$\widehat{F}_n(x) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i \le x), \quad \widehat{G}_m(x) = m^{-1}\sum_{i=1}^{m} \mathbf{1}(Y_i + \widehat{\delta}_N \le x), \quad x \in \mathbb{R}, \quad \text{(A.3)}$$

Then

$$X_1^*, \ldots, X_{n^*}^* \overset{i.i.d.}{\sim} \widehat{F}_n, \quad \text{(A.4)}$$

$$Y_1^*, \ldots, Y_{m^*}^* \overset{i.i.d.}{\sim} \widehat{G}_m. \quad \text{(A.5)}$$

Recall that the definition of the bootstrap signed rank statistic prior to centering is given by

$$\widetilde{W}_{N^*}^* = \binom{N^*}{2}^{-1} \sum_{i=1}^{n^*}\sum_{j=1}^{m^*} R_{ij}^* \mathbf{1}(X_i^* - Y_j^* > 0),$$

where R_{ij}^* denotes the rank of $|X_i^* - Y_j^*|$ among the bootstrap values $|X_i^* - Y_j^*|$, $i = 1, \ldots, n^*$, $j = 1, \ldots, m^*$. Further, define the two-sample bootstrap U-statistic as

$$U_{N^*}^* = \binom{n^*}{2}^{-1}\binom{m^*}{2}^{-1} \sum_{1\le i<i'\le n^*, 1\le j<j'\le m^*} \mathbf{1}(X_i^* - Y_j^* + X_{i'}^* - Y_{j'}^* > 0),$$

and the one-sample bootstrap U-statistic

$$\widetilde{U}_{N^*} = \binom{N^*}{2}^{-1} \sum_{(i-1)m^*+j<(i'-1)m^*+j'} \mathbf{1}(X_i^* - Y_j^* + X_{i'}^* - Y_{j'}^* > 0).$$

Note that for the expectation of $U_{N^*}^*$ under P^* we have

$$\theta_N^* = E^*(U_{N^*}^*) = P^*(X_1^* - Y_1^* > -(X_2^* - Y_2^*)).$$

Hence, we expect that the conditional distribution of $\sqrt{n^*+m^*}(U_{N^*}^* - \theta_{N^*}^*)$ converges a.s. to the same limit distribution as $\sqrt{n+m}(U_N - \theta(F,G))$. However, in the definition of $T_{N^*}^*$ we used the centering term

$$\widehat{C}_N = \binom{N}{2}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{1}(\widehat{D}_{ij} > 0)\widehat{R}_{ij}.$$

The following lemma shows that the difference between the correct centering term θ_N^* and $\widehat{C}_N$ is of the order $o_P(N^{-1/2})$ which ensures that we can use $\widehat{C}_N$ instead of θ_N^*.

LEMMA A.2. *Assume $n/m \to \lambda \in (0,\infty)$. Then*

(i) $\sqrt{n+m}(\theta_N^* - \widehat{C}_N) = o(1)$, *as* $N \to \infty$.

(ii) $\sqrt{n^*+m^*}T_{N^*}^* = \sqrt{n^*+m^*}(U_{N^*}^* - E^*(U_{N^*}^*)) + o_P(1) + o_{P^*}(1)$, *as* $N, N^* \to \infty$.

PROOF. Note that P^* is given by $P^*((X_1^*, Y_1^*) = (X_i, Y_j + \widehat{\delta}_N)) = (nm)^{-1}$, $i = 1,\dots,n$, $j = 1,\dots,m$. Hence

$$\theta_N^* = \frac{1}{n^2m^2} \sum_{i,i'=1}^{n} \sum_{j,j'=1}^{m} U_{ii'jj'}$$

with

$$U_{ii'jj'} = \mathbf{1}(X_i - Y_j + X_{i'} - Y_{j'} - 2\widehat{\delta}_N > 0)$$

is a two-sample V statistic with kernel

$$h(x_1, x_2, y_1, y_2) = \mathbf{1}(x_1 + x_2 - y_1 - y_2 > 0).$$

Let $\widehat{U}_{n,m} = \frac{1}{n(n-1)m(m-1)} \sum_{i\neq i'} \sum_{j\neq j'} U_{ii'jj'}$ denote the corresponding two-sample U-statistic. Recall that U- and V- statistics are known to be equivalent. Thus it suffices to show that $\sqrt{n+m}(\widehat{U}_{n,m} - \widehat{C}_N) = o_P(1)$. Note that $n^2m^2\widehat{U}_{n,m} - \theta_N^*$ is given by

$$\left(\sum_{i=i'} \sum_{j,j'} + \sum_{i,i'} \sum_{j=j'} + \left(\frac{1}{(n-1)(m-1)} + \frac{1}{n-1} + \frac{1}{m-1} \right) \sum_{i\neq i'} \sum_{j\neq j'} \right) U_{ii'jj'}.$$

All sums have not more than $O(n^3)$ terms. By boundedness of the kernel and the fact that $m = O(n)$ we have $n(\widehat{U}_{n,m} - \theta_N^*) = O(1)$, which verifies (i). Finally, note that the decomposition (A.1), Lemma A.1 with $\mu = P^*$, and the result just shown yield

$$\begin{aligned}
\sqrt{n^*+m^*}\, T_{N^*}^* &= \sqrt{n^*+m^*}\{\widetilde{W}_{N^*}^* - \widehat{C}_N\} \\
&= \sqrt{n^*+m^*}\{\widetilde{U}_{N^*} - \widehat{C}_N\} + o_{P^*}(1) \\
&= \sqrt{n^*+m^*}\{\widetilde{U}_{N^*} - \theta_N^*\} + o_{P^*}(1) + o(1) \\
&= \sqrt{n^*+m^*}\{U_{N^*}^* - \theta_N^*\} + o_{P^*}(1) + o(1) \\
&= \sqrt{n^*+m^*}\{U_{N^*}^* - E^*(U_{N^*}^*)\} + o_{P^*}(1) + o(1),
\end{aligned}$$

as $N \to \infty$ and $N^* \to \infty$, P^*-a.s. □

The following fact is used in the proof of Theorem 3.3, but it is also interesting in its own right.

LEMMA A.3. *We have $\widehat{G}_m(x) \to G(x)$, as $m \to \infty$, almost surely, for all $x \in \mathbb{R}$.*

PROOF. For convenience, we give an explicit proof avoiding abstract arguments. In what follows, we refer to the probability measure P. Fix $x \in \mathbb{R}$. By the Glivenko-Cantelli theorem, it suffices to show that

$$\left| m^{-1} \sum_{i=1}^{m} \mathbf{1}_{(-\infty, x+\widehat{\delta}_N]}(Y_i) - \mathbf{1}_{(-\infty, x+\delta]}(Y_i) \right| \overset{a.s.}{\to} 0,$$

as $m \to \infty$. The above expression is, of course, bounded by

$$B_m = m^{-1} \sum_{i=1}^{m} \mathbf{1}_{A_N}(Y_i),$$

where $A_N = [x + \widehat{\delta}_N, x + \delta) \cup [x + \delta, x + \widehat{\delta}_N)$. By Theorem 3.2 there exists a set Ω_0 with $P(\Omega_0) = 1$ and $\widehat{\delta}_N(\omega) \to \delta$, $N \to \infty$, for all $\omega \in \Omega_0$. Let $\varepsilon > 0$ and choose $N_0(\omega)$ such that for $N \geq N_0(\omega)$ we have $|\widehat{\delta}_N(\omega) - \delta| < \varepsilon$. Now for $\omega \in \Omega_0$ and $N \geq N_0(\omega)$ we have $[x + \delta, x + \delta + (\widehat{\delta}_N(\omega) - \delta)) \subset [x + \delta, x + \delta + \varepsilon)$ and $[x + \widehat{\delta}_N, x + \delta) \subset [x + \delta - \varepsilon, x + \delta)$. Thus, $A_N(\omega) \subset [x + \delta - \varepsilon, x + \delta + \varepsilon]$. Consequently,

$$m^{-1} \sum_{i=1}^{m} \mathbf{1}_{A_N(\omega)}(Y_i(\omega)) \leq m^{-1} \sum_{i=1}^{m} \mathbf{1}_{[x+\delta-\varepsilon, x+\delta+\varepsilon]}(Y_i(\omega)),$$

and there exists a set Ω_1 with $P(\Omega_1) = 1$ such that for all $\omega \in \Omega_0 \cap \Omega_1$ (an a.s. event) the right side of the above display converges to $G(x+\delta+\varepsilon) - G(x+\delta-\varepsilon) \to 0$, as $\varepsilon \to 0$. □

We are now in a position to prove Theorem 3.3.

PROOF OF THEOREM 3.3. By Lemma A.2 it suffices to verify that $P-$a.s. under the conditional bootstrap distribution P^*

$$\sqrt{n^* + m^*}(U^*_{N^*} - \theta^*_N) \overset{d}{\to} N(0, \eta^2(F, G)),$$

as $N, N^* \to \infty$. Since $U^*_{N^*}$ is a non-degenerate U-statistic, under the conditional bootstrap law P^* given the sample $\{X_i, Y_j : i = 1, \ldots, n, j = 1, \ldots, m\}$, we have under P^*, as $N^* \to \infty$,

$$\sqrt{n^* + m^*}(U^*_{N^*} - \theta^*_N) \overset{d}{\to} N(0, \widehat{\eta}^2_N),$$

where $\widehat{\eta}^2_N = \eta^2(\widehat{F}_n, \widehat{G}_m)$, i.e.,

$$\widehat{\eta}^2_N = 4\lambda^{-1}\widehat{\zeta}_{N,01} + 4(1-\lambda)^{-1}\widehat{\zeta}_{N,10}$$

with $\widehat{\zeta}_{N,01} = \zeta_{01}(\widehat{F}_n, \widehat{G}_m)$ and $\widehat{\zeta}_{N,10} = \zeta_{10}(\widehat{F}_n, \widehat{G}_m)$. Thus, it remains to show that the asymptotic variance, $\widehat{\eta}^2_N$, converges P-a.s. to $\eta^2(F, G)$. Since $\|\widehat{F}_n - F\|_\infty \to 0$

a.s., as $n \to \infty$, and Lemma A.3 yields $\|\widehat{G}_m - G\|_\infty \to 0$ a.s., as $m \to \infty$, we have to show that

$$\sup_z \left| \int (1 - \widehat{F}_n(y-z))\, d\widehat{G}(y) - \int (1 - F(y-z))\, dG(y) \right|,$$

if $n, m \to \infty$, P-a.s. Recall that the Skorohod space $D[0,1]$ consists of all right-continuous functions $[0,1] \to \mathbb{R}$ with existing left limits. It is known that the functional $\tau : (D[0,1])^2 \to \mathbb{R}$ given by

$$\tau(z, G) = \int z\, dG, \qquad (z, G) \in (D[0,1])^2,$$

is continuous in G with respect to the uniform topology, uniformly over those $z \in D[0,1]$ with $|z| \le K$ and $\int |dz| \le K$ for some constant $K > 0$. Here $\int |dz|$ denotes the total variation semi-norm of z. From this fact it is straightforward to conclude that

$$\widehat{\zeta}_{N,01} \overset{P-a.s.}{\to} \zeta_{01}(F, G), \qquad \widehat{\zeta}_{N,10} \overset{P-a.s.}{\to} \zeta_{10}(F, G),$$

as $n, m \to \infty$, implying the fact that $\widehat{\eta}^2 \to \eta^2$, as $m, n \to \infty$. Therefore, we obtain P-a.s.

$$T^*_{N^*} \overset{d}{\to} N(0, \eta^2(F, G)),$$

as $N, N^* \to \infty$, provided $n^*/(n^* + m^*)$ converges to the same limit as $n/(n+m)$, in order to ensure that the asymptotic variances coincide. This completes the proof. □

A.2. Proof of the multi-sample bootstrap central limit theorem. Finally we give a sketch of the proof for the multi-sample setting.

Proof of Theorem 4.1. We show (ii). Assertion (i) follows using the same arguments. By the Cramer-Wold device we have to verify that for all $(\rho_{kl})_{(k,l)\in\mathcal{H}_0}$ with $\prod_{k,l} |\rho_{kl}| \neq 0$,

$$\sum_{(k,l)\in\mathcal{H}_0} \rho_{kl} T^*_{kl} \overset{d}{\to} N\left(0, \sum_{(k,l)\in\mathcal{H}_0} \rho^2_{kl} \eta^2(F_k, F_l)\right), \tag{A.6}$$

holds, if $n_k, n^*_k, n_l, n^*_l \to \infty$. Under the constraints given in the theorem, Lemma A.2 yields

$$T^*_{kl} = \sqrt{n^*_k + n^*_l}(U^{(k,l)*}_{N^*_{kl}} - \theta^*_{kl}) + o_P(1) + o_{P^*}(1).$$

where

$$\begin{aligned} U^{(k,l)*}_{N^*_{kl}} &= \binom{n^*_k}{2}^{-1} \binom{n^*_l}{2}^{-1} \sum_{i<i'} \sum_{j<j'} \mathbf{1}(X^*_{ki} - X^*_{lj} + X^*_{ki'} - X^*_{lj'} > 0), \\ \theta^*_{kl} &= E^*(U^{(k,l)*}_{N^*_{kl}}) \end{aligned}$$

for all $(k,l) \in \mathcal{H}_0$. Hence, since $\mathcal{H}_0$ is a finite set,

$$\sum_{k,l} \rho_{kl} T^*_{kl} = \sum_{k,l} \rho_{kl} \sqrt{n^*_k + n^*_l}(U^{(k,l)*}_{N^*_{kl}} - \theta^*_{kl}) + o_P(1)$$

A finite set of multi-sample U-statistics is jointly asymptotically normal, see Lehmann (1963) or Randles and Wolfe (1979, Theorem 3.6.9). Therefore, under the conditional bootstrap distribution, $(\sqrt{n^*_k + n^*_l}(U^{(k,l)*}_{N^*_{kl}} - \theta^*_{kl}) : (k,l) \in \mathcal{H}_0)$ is asymptotically normal with asymptotic variances $\widehat{\eta}^2 = \eta^2(\widehat{F}_{n_k}, \widehat{G}_{n_l})$. Lemma A.3 applied to the sample $X_{l1} + \widehat{\delta}_{kl}, \ldots, X_{ln_l} + \widehat{\delta}_{kl}$ yields a.s. convergence to $\eta^2(F_k, F_l)$, as $n_k, n_l \to \infty$. It remains to study the asymptotic covariances. Note that for $l \neq m$, we have

$$E^*(T^*_{kl} T^*_{km}) = \sqrt{n^*_k + n^*_l}\sqrt{n^*_k + n^*_m} E^*(U^{(k,l)*}_{N_{kl}} - \theta^*_{kl})(U^{(k,m)*}_{N^*_{km}} - \theta^*_{km}) + o_P(1).$$

Further,

$$E^*(U^{(k,l)*}_{N_{kl}} - \theta^*_{kl})(U^{(k,m)*}_{N^*_{km}} - \theta^*_{km}) =$$
$$\frac{1}{\binom{n^*_k}{2}^2 \binom{n^*_l}{2}\binom{n^*_m}{2}} \sum_{1 \le i < i' \le n^*_k, 1 \le j < j' \le n^*_l} \sum_{1 \le r < r' \le n^*_k, 1 \le s < s' \le n^*_m} E^*(V_{klii'jj'} V_{kmrr'ss'}),$$

where

$$\begin{aligned} V_{klii'jj'} &= \mathbf{1}(X^*_{ki} - X^*_{lj} + X^*_{ki'} - X^*_{lj'} > 0) - \theta^*_{kl} \\ V_{kmrr'ss'} &= \mathbf{1}(X^*_{kr} - X^*_{ms} + X^*_{kr'} - X^*_{ms'} > 0) - \theta^*_{km} \end{aligned}$$

$V_{klii'jj'}$ and $V_{kmrr'ss'}$ are independent under P^*, if $i \neq r$ and $i \neq r'$ and $i' \neq r$ and $i' \neq r'$ yielding $E^*(V_{klii'jj'} V_{kmrr'ss'}) = 0$. The above argument applies to $\binom{n^*_k}{2}\binom{n^*_k - 2}{2}\binom{n^*_l}{2}\binom{n^*_m}{2}$ terms. Therefore,

$$E^*(T^*_{kl} T^*_{km}) = o(1) + o_P(1). \tag{A.7}$$

Let $h(x_1, x_2, y_1, y_2) = \mathbf{1}(x_1 - y_1 + x_2 - y_1 < 0)$ denote the kernel function inducing the two-sample U-statistics. Notice that the four-sample U-statistic

$$\frac{1}{\binom{n^*_k}{2}\binom{n^*_l}{2}\binom{n^*_k}{2}\binom{n^*_m}{2}} \sum_{i,i'} \sum_{j,j'} \sum_{r,r'} \sum_{s,s'} [\rho_{kl} h(X_{ki}, X_{ki'}, Y_{lj}, Y_{lj'}) + \rho_{km} h(X_{kr}, X_{kr'}, Y_{ms}, Y_{ms'})]$$

equals the linear combination of two-sample U-statistics

$$\rho_{kl} \frac{1}{\binom{n^*_k}{2}\binom{n^*_l}{2}} \sum_{i,i'} \sum_{j,j'} h(X_{ki}, X_{ki'}, Y_{lj}, Y_{lj'}) + \rho_{km} \frac{1}{\binom{n^*_k}{2}\binom{n^*_m}{2}} \sum_{r,r'} \sum_{s,s'} h(X_{ki}, X_{ki'}, Y_{mj}, Y_{mj'}),$$

since the first term of each summand does not depend on r, r', s, s', and the second one does not depend on i, i', j, j'. Hence, we conclude that $\sum_{(k,l)\in\mathcal{H}_0} \rho_{kl} T_{kl}$ is asymptotically equivalent to a multi-sample U-statistic, the summands being asymptotically uncorrelated using (A.7). Thus, we conclude that $\sum_{k,l\in\mathcal{H}_0} \rho_{kl} T^*_{kl}$ is asymptotically normal with asymptotic variance given by $\sum_{(k,l)\in\mathcal{H}_0} \rho_{kl} \eta(F_k, F_l)^2$, that is, (A.6) holds. □

Ansgar Steland
Faculty of Mathematics,
Computer Science and Natural Science
Institute of Statistics
RWTH Aachen University
D-52065 Aachen, Germany
E-mail: steland@stochastik.rwth-aachen.de

Appaswamy R. Padmanabhan
Department of Mathematics
and Statistics
Monash University
Clayton, VIC 3800
Australia
E-mail: padmanabhan6001@hotmail.com

Muhammad Akram
Faculty of Medicine, Nursing
and Health Sciences
Department of Epidemiology
and Preventive Medicine
Monash University
Clayton, VIC 3800, Australia
E-mail: Muhammad.Akram@monash.edu

Paper received December 2009; revised November 2010.